A Genomic Perspective on Protein Families

Roman L. Tatusov, Eugene V. Koonin,* David J. Lipman

In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

The release in 1995 of the complete genome sequence of the bacterium *Haemophilus influenzae* (*1*), followed within the next 1.5 years by four more bacterial genomes (*2*), one archaeal genome (*3*), and one genome of a unicellular eukaryote (*4*), marked the advent of a new age in biology. The hallmark of this era is that comparisons between complete genomes are becoming an indispensable component of our understanding of a variety of biological phenomena. The number of sequenced genomes is expected to grow exponentially for at least the next few years, and conceivably, their impact on biology will further increase (*5*).

Knowing the inventory of conserved genes responsible for housekeeping functions and understanding the differences in the genetic basis of these functions in different phylogenetic lineages is central to understanding life itself, at least at the level of a single cell. Complete sequences are indispensable for achieving this goal because they hold the only type of information that can be used to delineate the complete network of relationships between genes from different genomes. Furthermore, only with complete genome sequences is it possible to ascertain that a particular protein implicated in an essential function is not encoded in a given genome. Accordingly, an alternative protein for the respective function should be sought among the functionally unassigned gene products (*6*). With multiple genome sequences, it is possible to delineate protein families that are highly conserved in one domain of life but are missing in the others. Such information may be critically important: For example, the families that are conserved among bacteria but are missing in eukaryotes comprise the pool of potential targets for broad-spectrum antibiotics.

The knowledge of all of the gene sequences from multiple complete genomes redefines the problem of gene classification. It becomes feasible to replace the more or less arbitrary clustering of genes by similarity with a complete, consistent system in which the groups are likely to have evolved from a single ancestral gene. Such a natural classification of genes will provide a framework for evolutionary studies and for rapid, largely automatic functional annotation of newly sequenced genomes. This framework will evolve and improve with increasing coverage of the diversity of life forms with complete genome sequences. It is critical to have this system in place while the number of completed genomes is still small and each family can be explored individually. Here we describe a prototype of a natural system of gene families from complete genomes.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. Orthologs are genes in different species that evolved from a common ancestral gene by speciation; by contrast, paralogs are genes related by duplication within a genome (*7*). Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original one. Thus, identification of orthologs is critical for reliable prediction of gene functions in newly sequenced genomes. It is equally important for phylogenetic analysis because interpretable phylogenetic trees generally can be constructed only within sets of orthologs (*8*). A complete list of orthologs also is a prerequisite for any meaningful comparison of genome organization (*9*).

A naïve operational definition would simply maintain that for a given gene from one genome, the gene from another genome with the highest sequence similarity is the ortholog. Given the complete genome sequences, this straightforward approach often gives credible results, especially when the compared species are not too distant phylogenetically (*9*). At larger phylogenetic distances, however, the situation becomes more complicated. If gene duplications occurred in each of the given two clades subsequent to their divergence, only a many-to-many relationship will adequately describe orthologs, and accordingly, detection of the highest similarity will not result in the identification of the complete set of orthologs. In addition, when the best hit is not highly significant statistically, which is common in the case of phylogenetically distant relationships (*10*), it simply may be spurious. On the other hand, attempts to apply a restrictive similarity cutoff are likely to result in a number of orthologs being missed.

Given the existence of one-to-many and many-to-many orthologous relationships, we redefined the task of identifying orthologs as the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages. In other words, any two proteins from different lineages that belong to the same COG are orthologs. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

In order to delineate the COGs, all pairwise sequence comparisons among the 17,967 proteins encoded in the seven complete genomes were performed (*11*), and for each protein, the best hit (BeT) in each of the other genomes was detected. The identification of COGs was based on consistent patterns in the graph of BeTs. The simplest and most important of such patterns is a triangle, which typically consists of orthologs (Fig. 1A). Indeed, if a gene from one of the compared genomes has BeTs in two other genomes, it is highly unlikely that the respective genes are also BeTs for one another unless they are bona fide orthologs (*12*). The consistency between BeTs resulting in triangles does not depend on the absolute level of similarity between the compared proteins and thus allows the detection of orthologs among both slowly and quickly evolving genes. This approach is most likely to be informative when the

The authors are with the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

*To whom requests for reprints should be addressed. E-mail: koonin@ncbi.nlm.nih.gov

EXHIBIT 2

BEST AVAILABLE COPY

BeTs forming a triangle come from widely different lineages. Accordingly, only five major, phylogenetically distant clades were used as independent contributors to COGs: Gram-negative bacteria (*Escherichia coli* and *H. influenzae*), Gram-positive bacteria (*Mycoplasma genitalium* and *M. pneumoniae*), Cyanobacteria (*Synechocystis* sp.), Archaea (Euryarchaeota) (*Methanococcus jannaschii*), and Eukarya (Fungi) (*Saccharomyces cerevisiae*) (*13*).

The procedure used to derive COGs included finding all triangles formed by BeTs between the five major clades and merging those triangles that had a common side until no new ones could be joined. A triangle is an elementary, minimal COG (Fig. 1A). The groups produced by merging adjacent triangles include orthologs from different lineages and, in many cases, paralogs from the same lineage (Fig. 1, B and C). Because of the existence of paralogs, the BeTs that form the triangles are not necessarily symmetrical: For example, in the COG shown in Fig. 1C, the same *M. genitalium* protein, MG249, is the BeT for four paralogous σ subunits of *E. coli* RNA polymerase, but only for one of them, RpoD, is the relationship symmetrical.

Most of the clusters derived by the above procedure meet the definition of a COG, that is, all of the proteins from the different lineages in the same cluster are likely to be orthologs. There are, however, several reasons why, in certain cases, COGs may be lumped together. Proteins may contain two or more distinct regions, each of which belongs to a different conserved family; usually such proteins are loosely referred to as multidomain (*14*). Each of the clusters was inspected for the presence of multidomain proteins, individual domains were isolated (*15*), and a second iteration of the sequence comparison was performed with the resulting database of domains. Some of the COGs may include proteins from different lineages that are paralogs rather than orthologs, primarily because of differential gene loss in the major phylogenetic lineages. When one gene in a pair of paralogs is lost in one lineage but not in the others, two COGs that should have been distinct may be artificially joined. Therefore, the level of sequence similarity between the members of each cluster was analyzed, and clusters that seemed to contain two or more COGs were split.

Phylogenetic and Functional Patterns in COGs

The described analysis resulted in 710 apparent COGs. This set appears to be essentially complete as far as orthologous relationships are concerned. Indeed, when the portion of the database of proteins from complete genomes not included in the COGs was clustered by sequence similarity (*16*), only 10 groups were identified, which, upon careful inspection of the alignments, were considered likely to constitute additional COGs missed originally. These groups were incorporated, producing the final collection of 720 COGs, including 6814 proteins and distinct domains of multidomain proteins (6646 distinct gene products, or 37% of the total number of genes in the seven complete genomes) (*17*).

Most of the COGs are relatively small groups of proteins. One-third of the COGs (240 COGs with 1406 proteins) contain one representative of each of the included species (no paralogs), and 192 more COGs include paralogs from only one species, most frequently yeast (87 COGs). The mean number of proteins per COG increases with increasing number of genes in a genome, from 1.2 for *M. genitalium* to 2.9 for yeast. A notable aspect of many COGs is the differential behavior of paralogs. It is typical that one of the paralogs, for example, in yeast, shows consistently higher similarity to the orthologs in all or most of the other species (Fig. 1, B and C). For numerous yeast paralogs, particularly components of the translation apparatus, the underlying cause is obvious: the gene whose product is most similar to the bacterial orthologs is of mitochondrial origin (Fig. 1B). A more common explanation for the asymmetry of the relationships in the COGs, however, is that the highly conserved paralog has retained the original function, whereas the functions of the less conserved paralogs have changed in the course of evolution. In the already considered example (Fig. 1C), the symmetrical component of the graph (solid lines) delineates the conserved function of the σ70 subunit of the RNA polymerase (*E. coli* RpoD), which is required for the transcription of the bulk of bacterial genes, whereas the asymmetrical BeTs (broken lines) are observed for σ subunits (*E. coli* RpoH, RpoS, and FliA) involved in the transcription of specialized gene subsets (*18*). This phenomenon appears to be widespread, as we found 549 proteins in 302

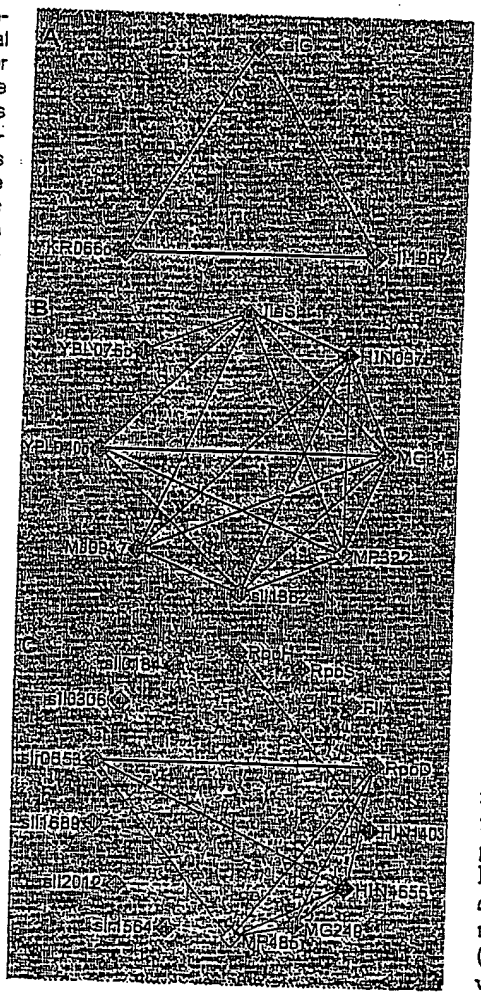

Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two yeast paralogs. Origin of the proteins: IleS, *E. coli*; HIN0378, *H. influenzae*; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; and YBL076c and YPL040c, *S. cerevisiae*. Note the adjacent triangles with a common side, for example, IleS-MG345-MJ0947 and sll1362-MG345-MJ1362. YPL040c is the yeast mitochondrial isoleucyl-tRNA synthetase; the bacterial orthologs and that from *M. jannaschii* are the BeTs for this yeast protein, but the reverse is true only of the bacterial proteins (symmetrical BeTs). Conversely, for YBL076c, which is the yeast cytoplasmic isoleucyl-tRNA synthetase, the *M. jannaschii* ortholog is a symmetrical BeT, whereas the bacterial BeTs are asymmetrical. (**C**) A complex COG with multiple paralogs. Origin of the proteins: RpoH, RpoS, RpoD, and FliA, *E. coli*; HIN1403 and HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, and slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, and MG249 are major sigma factors (σ70), whose function is universal in bacteria; note the fully symmetrical relationships between these proteins. The other proteins are specialized sigma factors whose radiation from the ancestral family apparently was accompanied by modification of the function and involved accelerated evolution; note the asymmetrical BeTs.

BEST AVAILABLE COPY

COGs whose corresponding paralogs showed consistently lower similarity to other members of the COG. One may think of the rapidly evolving paralogs as progenitors of new families emerging from within the conserved ones. The COGs will be an important resource in a systematic survey of the functional diversification of paralogs in conserved gene families.

There are several large clusters in the current collection with complex relationships between members. Two of these, namely the adenosine triphosphatase (ATPase) components of ABC transporters and histidine kinases, each include over 100 members. It is likely that subsequent detailed analysis of these large groups (for example, by phylogenetic tree methods) will result in their split into several distinct COGs, especially when more genomes are available. On a more general note, COGs do not supplant traditional methods of phylogenetic analysis but rather provide the appropriate starting material for these methods, in particular for a systematic analysis of phylogenetic tree topology.

Figure 2 shows the breakdown of the COGs by broadly defined function (*19*) and by species (*20*). For the majority of the COGs, the protein function is either known from direct experiments, mainly in *E. coli* or yeast, or can be confidently inferred on the basis of significant sequence similarity to functionally characterized proteins from other species. It has to be emphasized that construction of the COGs includes automatic prediction of the function for numerous genes, particularly from the poorly characterized genomes such as *M. jannaschii*. There is, however, a substantial fraction of the COGs (14%) for which only general functional prediction, typically of biochemical activity, but not the actual cellular role could be made, and for another 5%, there was no functional clue (Fig. 3). Each of the COGs includes proteins from at least three major clades whose divergence time is estimated to be over a billion years (*21*), that is, they all are ancient, conserved families with important, if not necessarily essential, cellular functions. Therefore, the proteins belonging to the "mysterious" COGs are good candidates for directed experimental studies.

The distribution of proteins from different species in the COGs shows several trends (Fig. 2), although the bias in the current collection of complete genomes (in particular, because three lineages are required to form a COG, all COGs had to have a bacterial member) must be taken into account when interpreting these comparisons. The fraction of proteins belonging to COGs is greatest in the nearly minimal genomes of mycoplasmas (70% for *M. genitalium*) and much lower in the larger genomes of *E. coli* and yeast (40% and 26%, respectively), which indeed is the tendency expected of conserved families presumably associated with cellular housekeeping functions. The genes of the pathogenic bacteria (*H. influenzae* and two mycoplasmas) are essentially subsets of the two larger bacterial gene complements, *E. coli* and *Synechocystis* sp. The latter two species almost always cooccur in the COGs. The main cause of the observed congruency is likely to be the conservation of the core of ancestral bacterial genes in nonparasitic species from different major clades. Accordingly, the fact that proteins from the pathogenic bacteria are missing in many COGs most likely testifies to gene loss, which has been extensive even in this subset of highly conserved genes. The cooccurrence of *M. jannaschii* in a COG with *E. coli* or *Synechocystis* is measurably more frequent than that with yeast (Fig. 2). Such a distribution of the archaeal genes appears to be due primarily to the blending of bacterial-like and eukaryotic-like genes in the archaeal genomes (*10*), although the mentioned bias in the genome collection is also a factor.

The phylogenetic distribution of the COG members is distinct for different functional classes (Fig. 2). It is not unexpected that translation is the only category in which ubiquitous COGs are predominant. Another obvious trend is the absence of proteins from pathogenic bacteria (*H. influenzae* and, particularly, the mycoplasmas) in many COGs

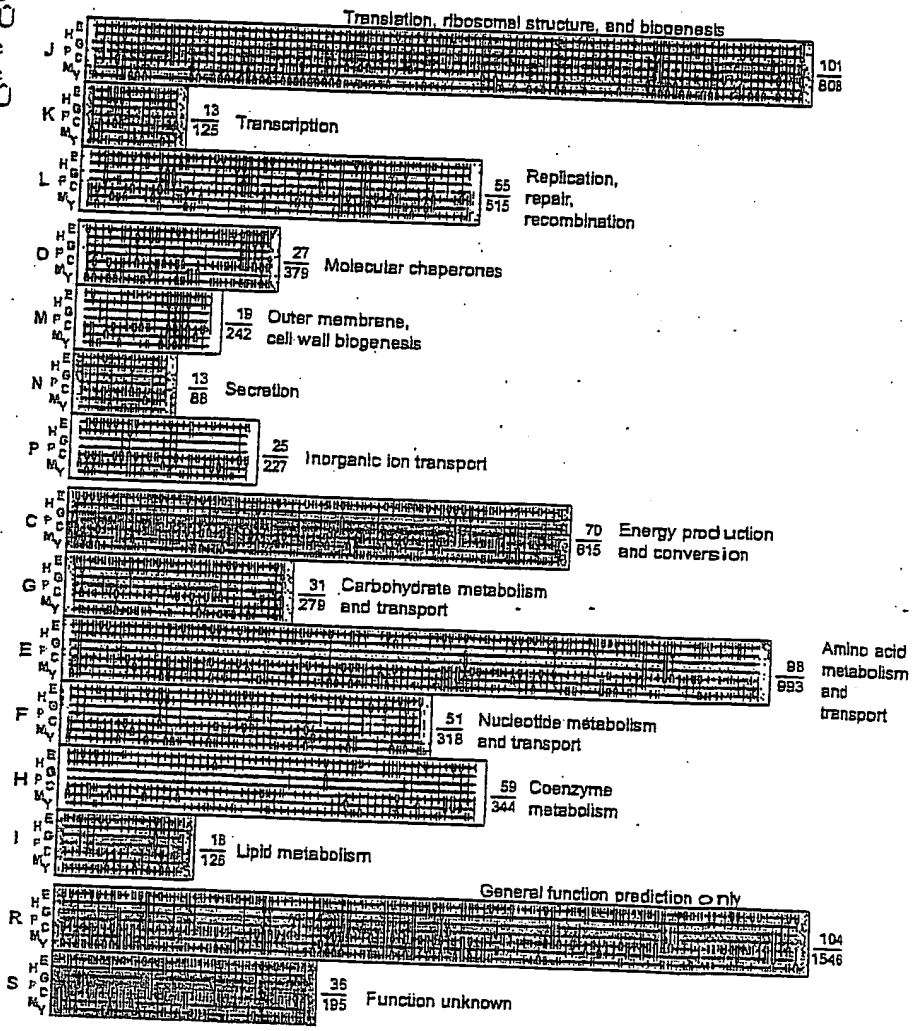

Fig. 2. A functional and phylogenetic breakdown of the COGs. E indicates *E. coli*; H, *H. influenzae*; G, *M. genitalium*; P, *M. pneumoniae*; C, *Synechocystis* sp.; M, *M. jannaschii*; and Y, *S. cerevisiae*. Each column shows a COG; a double streak indicates that two or more paralogs from the given species belong to the particular COG. The number of COGs (numerator) and the number of proteins in them (denominator) is indicated for each functional category. Capital letters in the leftmost field encode the functional categories (used in the COG IDs).

BEST AVAILABLE COPY

in each functional category other than translation and transcription, but especially in the metabolic functional classes. Conversely, the congruence between the two nonparasitic bacteria, *E. coli* and *Synechocystis* sp., holds for all functional classes (Fig. 2). Also apparent is the differential appearance of archaeal proteins that tend to group with yeast proteins in the translation and transcription classes (which, given the bias in the genome collection, results in ubiquitous COGs) but in all other functional classes are frequently found in COGs with bacterial proteins only.

The phylogenetic distribution of COG membership can be conveniently presented in terms of "phylogenetic patterns," which show the presence or absence of each analyzed species (Fig. 3). Of the 88 patterns that include at least three lineages (the definition of a COG), 36 were actually found. Missing were mostly patterns with only one of the two species of *Mycoplasma*, which was predictable because the gene complement of *M. genitalium* is essentially a subset of the *M. pneumoniae* complement (22). The remaining eight patterns that were never observed all include pathogenic bacteria without *E. coli*, which is the largest and most diverse of the available bacterial genomes. The two most abundant patterns could easily be predicted: all species ("ehgpcmy"), and all species except for the mycoplasmas ("eh__cmy"). What appears much less trivial is that these patterns together encompass only one-third of all COGs. This fact emphasizes the remarkable fluidity of genomes in evolution, revealed in spite of the fact that the analysis concentrated on ancient conserved families. Multiple solutions for the same important cellular function appear to be a rule rather than an exception, at least when phylogenetically distant species are considered (*10*, *23*). On the other hand, the eight most frequent patterns, which together account for 85% of the COGs, all include both *E. coli* and *Synechocystis*, emphasizing the congruency between these genomes.

The 114 ubiquitous COGs, most of them including components of the translation and transcription machinery, form the universal core of life. This set is more than twofold down from the bacterial "minimal set" consisting of 256 genes (*23*), but significant further erosion seems unlikely, given the broad spectrum of compared genomes.

The higher order distribution of the COGs by the three domains of life, with only 45% of the COGs including representatives of Bacteria, Archaea, and Eukarya, is another manifestation of the dynamics of gene families in evolution (Fig. 3). The picture is expected to become even more complex, and the fraction of three-domain COGs will probably drop, once archaeal-only, eukaryotic-only, and archaeal-and-eukaryotic COGs emerge with the accumulation of genome sequences.

The unusual, rare patterns are of particular interest, suggesting the possibility of unexpected findings. Each of the COGs with patterns that occur only once in our current collection (Table 1) should correspond to a unique function scattered over disconnected branches of the tree of life. Why such functions are conserved and are presumably important for survival in some but not other lineages is a challenge to be addressed experimentally. The principal evolutionary mechanisms that can be invoked to explain the emergence of these rare patterns are differential gene loss and horizontal transfer of genes. Some of the functions involved, for example, lipoate-protein ligase and glycyl–transfer ribonuclease (tRNA) synthetase, appear to be strictly essential, but in different species, they are performed by two distinct sets of orthologs unrelated to one another (*24*). Other functions, for example, thymidine phosphorylase and hexuronate dehydrogenases, may be dispensable under most conditions, and accordingly, differential gene loss is likely; it is remarkable, however, that these functions are preserved in the nearly minimal gene complements of the mycoplasmas. Two of the unique patterns, namely "__gpc_y" and "_hgp__y," might have evolved through horizontal transfer of typical eukaryotic genes into bacterial genomes. The latter pattern is of particular interest as it involves the choline kinase gene common to a number of bacterial pathogens and implicated in pathogenicity (*25*). Two of the COGs with unique patterns, "h__c_y" and "e_gp_my," include highly conserved but uncharacterized proteins whose functions could be predicted only by detailed analysis of conserved protein motifs (Table 1). These examples demonstrate the potential for protein function prediction inherent in the construction of the COGs themselves.

The sampling of genomes we compared is small and biased, and when a more complete set is available, the distribution of COGs by phylogenetic patterns is likely to change significantly; for example, many patterns that are currently rare may become common when larger genomes from the Gram-positive bacterial lineage (such as *Bacillus subtilis*) become available. Nevertheless, we believe that the language of phylogenetic patterns will become even more useful for the description of relationships between multiple genomes.

Connecting and Expanding the COGs

Ancient families of paralogs that span a broad range of taxa are well known (*26*). Accordingly, a number of COGs are related to each other and can be connected into superfamilies. In order to elucidate the superfamily structure of the COG collection, we used the recently developed PSI-BLAST (position-specific iterative BLAST) program, which combines BLAST search with profile analysis (*27*). Two COGs were considered connected if at least two of the proteins from the first COG hit members of the second COG in the PSI-BLAST search, and vice versa. Clustering by this criterion produced 58 superfamilies including 280 COGs.

Compared to COGs themselves, the superfamilies are a higher level of protein classification. Typically, they include conserved motifs that are determinants of a distinct biochemical activity, which, however, may be required for a variety of cellular functions. For example, the largest superfamily contains 53 COGs with 863 proteins, all of which contain conserved motifs typical of ATPases and GTPases but are involved in a broad range of processes from DNA replication to metabolite transport (*28*).

Superfamilies and their signature motifs

Bacteria+Eukarya +Archaea		Bacteria+Eukarya		Bacteria+Archaea		Bacteria only	
Pattern	COGs	Pattern	COGs	Pattern	COGs	Pattern	COGs
e___cmy	37					e_gpc__	5
eh___my	18	ehgp__y	5	ehgpcm_	15	eh_pc__	2
____cmy	13	e_gpc_y	2	e_gpcm_	4		
e____my	7	e__p__y	1	_h__m_	3		
__gpcmy	4	e_gp__y	1	eh_p_m_	2		
_h___my	2	eh_pc_y	1	ehgp_m_	2		
eh_p_my	2	_h__c_y	1	e_gp_m_	1		
ehgp_my	2	__gpc_y	1				
e_gpcmy	2	_hgp__y	1				
__gp_my	1						
e_gp_my	1						
eh_pcmy	1						
Sum	323	215		122		60	
COGs(%)	45	30		17		8	

Fig. 3. Phylogenetic patterns in COGs. Letter codes as in Fig. 2 (ignore case); an underline indicates absence of the respective species. Shading indicates the eight most frequent patterns.

will be useful in classifying proteins that have evolved to an extent that they can not be assigned to any COG but still retain a conserved motif. We sought to detect such proteins with distant, subtle similarity to COGs that might be encoded in the analyzed genomes. The PSI-BLAST analysis (*27*) detected "tails" of distantly related proteins (a total of 3686) for 321 COGs, increasing the total number of proteins connected to COGs to 10,332 (58% of the entire protein set from complete genomes).

Because apparent orthologs from at least three major clades were required to form a COG, there are potential new COGs hidden among the results of the comparison of protein sequences from complete genomes (*11*). Clustering by sequence similarity the proteins not included in COGs (*14*) resulted in 443 groups with members from two clades. Predictably, the greatest number, 204, were from the cyanobacterial and Gram-negative clades, followed by 67 groups combining yeast and *M. jannaschii*. Many of these groups are likely to become COGs once additional genomes are included in the analysis.

Prediction of Protein Functions with the COG System

The COG system allows automatic functional and phylogenetic annotation of genes and gene sets (*29*). As in the procedure used for the construction of the COGs, the criterion for adding likely orthologs from other genomes to the COGs is based on the consistency between the observed relationships. A protein is compared to the database of protein sequences from complete genomes (*11*) and is included in a COG if at least two BeTs fall into it. Given that the COGs were constructed from proteins encoded in complete genomes, it is not a requirement that newly included proteins also originate from a complete genome. Indeed, while the unsequenced portion of a genome may encode proteins with the highest similarity to those included in COGs, the BeTs will not change for the products of already sequenced genes.

As a demonstration of the principle coupled with additional characterization of the COGs themselves, the sequences of proteins with known three-dimensional structures from the PDB database (*30*) were compared to the protein sequences encoded in complete genomes. The "two BeT" procedure resulted in proteins with known three-dimensional structure being included in 183 COGs, of which one was shown to be a false positive by subsequent alignment analysis. Thus, structural information could be inferred for at least 25% of the COGs. In most cases, the structurally characterized protein (from *E. coli* or yeast) actually belongs to a COG or is a closely related homolog of the proteins forming a COG.

Some of the predictions, however, provide significant functional and structural inferences. Of particular interest are (i) the possibility of modeling the nuclease domain of polyadenylate cleavage factors

Table 1. Unique phylogenetic patterns among COGs. The pattern designations are as in Fig. 3; each COG ID includes a letter indicating the functional category, to which the constituent proteins belong (Fig. 2).

Pattern and COG ID	Proteins	Activity or function	Comment
e_gp_m_ COG0213F	DeoA-MG051-MP090-MJ0667	Thymidine phosphorylase; salvage of deoxypyrimidines	Nonessential gene in *E. coli*; apparent orthologs found in other Gram-positive bacteria and in humans (*35*).
e__p__y COG0246G	MtlD, UxaB, UxuB, YdfI, YeiQ-MP190-YEL070w, YNR073c	Mannitol-1-phosphate and other hexuronate dehydrogenases; hexuronate catabolism	Nonessential genes in *E. coli*; accessory reactions of carbohydrate metabolism (*36*).
e_gp__y COG0095H	LplA-MG270-MP450-(sll0809)-YJL046w	Lipoate-protein ligase A; ligation of lipoate to apoproteins of pyruvate dehydrogenase and other lipoate-dependent enzymes	There are two unrelated classes of lipoate-protein ligases; *E. coli* and yeast encode both forms; *H. influenzae* and *Synechocystis* sp. encode the B form (included in a separate COG); sll0809 is a distant homolog of the A form (*37*), which was not automatically included in the COG but was detected with PSI-BLAST.
eh_pc_y COG0604R	AdhC + 18 *E. coli* proteins–MP278-sll0990, slr1192-YBR046c + 19 yeast proteins	Alcohol dehydrogenase class III and related Fe-S dehydrogenases; various catabolic pathways	Highly conserved protein family distinct from other Fe-S oxidoreductases.
_h__c_y COG0678R	HIN1693_1-sll1621-YLR109w	Glutaredoxin-like membrane protein (prediction)	The *H. influenzae* protein contains an additional thioredoxin-like domain.
__gpc_y COG0631R	MG108-MP586-sll1771-sll1033-sll0602-YDL006w + 6 yeast proteins	Protein serine and threonine phosphatase	Serine and threonine protein phosphatases are abundant in eukaryotes but not in bacteria (*38*).
__gp_my COG0423J	MG251-MP483-MJ0228-YPR081c, YBR121c	Glycyl-tRNA synthetase (eukaryotic and Gram-positive type)	Gram-negative bacteria and *Synechocystis* encode a distinct glycyl-tRNA that appears to be unrelated to the eukaryotic and Gram-positive type; the closest relative of this COG in *E. coli* and *H. influenzae* is prolyl-tRNA synthetase (*24*).
e_gp_my COG0622R	b2300-MG207, MP029-MJ0623, MJ0936-YHR012w	Phosphoesterase (prediction)	Highly conserved protein family that shares only modified catalytic motifs (detected by PSI-BLAST; $P \sim 0.004$) with other phosphoesterases, including protein phosphatases.
eh_pcmy COG0078E	ArgI, ArgF, YgeW-HIN0012-MP531-sll0902-MJ0881-YJL088w	Ornithine carbamoyltransferase; arginine biosynthesis	Amino acid metabolism appears to be completely missing in *M. genitalium*, but residual reactions may occur in *M. pneumoniae*.
_hgp_y COG0510M*	HIN0936-MG356, MP310-YDR147w, YLR133w	Choline kinase (prediction) involved in lipopolysaccharide biosynthesis	Enzyme common to several bacterial pathogens and eukaryotes; contributes to pathogenicity (*25*).

*This COG was added to the collection by cluster analysis.

(31) with the beta-lactamase structure, (ii) the presence of an acylphosphatase domain in hydrogenase expression factors, which form a highly conserved COG, and in a number of uncharacterized proteins, and (iii) the connection between a unique carbonic anhydrase and an acetyltransferase family (Table 2).

Probably the most important application of the COGs is functional characterization of newly sequenced genomes. In the preliminary analysis of the recently published genome of the major human bacterial pathogen *Helicobacter pylori* (32), 813 proteins (51% of the gene products) from this bacterium were included in 453 preexisting COGs and 143 new COGs (33). In spite of the fact that many *H. pylori* proteins are highly similar to homologs from *E. coli* and other bacteria and have been explored in detail (32), this analysis produced over 100 additional functional predictions (33).

Conclusions and Perspective

The COGs bring together the fields of comparative genomics and protein classification. Among the numerous possible approaches to protein classification, the COGs appear to be unique as a prototype of a natural system, which has as its basic unit a group of descendants of a single ancestral gene. Typically, such a group is associated with a conserved, specific function, so that the inclusion of a protein in a COG automatically entails functional prediction.

Each COG contains conserved genes from at least three phylogenetically distant clades and, accordingly, corresponds to an ancient conserved region (ACR). Previous analyses have indicated that the total number of distinct ACRs is likely to be less than 1000 (34). Thus, even with the limited number of complete genomes currently available for analysis, the COGs have already captured a substantial fraction of all existing highly conserved protein domains. With more genomes included in the system, the discovery of additional COGs should gradually level off, with the great majority of the ACRs encoded in the added genomes fitting into already known COGs.

With the forthcoming flood of genome sequences, a coherent framework for understanding these genomes from both the functional and evolutionary viewpoints is a must. We regard the current collection of

Table 2. Structural and functional predictions for uncharacterized proteins in COGs.

Phylogenetic pattern and COG ID*	Proteins in COG†	Activity and function	Homolog in PDB‡ •BeTs detected (no.) •Lowest *P* with a COG member	Comment
e_gpcmy COG0595R	PhnP, ElaC-2g-2p-5c-8m-YLR277c, YMR137c, YKR079c	Predicted Zn-dependent hydrolases	Beta-lactamase (1BMC) •2 •0.039	Activity is not known for any protein in this ubiquitous COG. Biochemical and genetic data indicate that YLR277c is involved in messenger RNA 3′-end processing (37), whereas YMR137c is DNA cross-link repair protein SNM1 (39). A motif including the Zn-coordinating histidines of beta-lactamase is conserved.
eh__cmy COG0607R	SseA, PspE, GlpE, YibN, YbbB, YnjE, YgaP-2h-5c-MJ0052-4y	Predicted sulfurtransferases	Rhodanese (1RHD, 2ORA, 1ORB) •2 •10^{-41}	The sulfurtransferase activity of SseA has been demonstrated (40), but the rest of the proteins in this COG have no known activity. PspE (phage shock protein), GlpE (uncharacterized protein involved in glycerol metabolism), and other small proteins correspond to one of the two rhodanese domains.
ehgpc_y COG0596R	PldB, MhpC, YcdJ, YnbC-HIN0065-MG020-MP132-6c-YNR064c, YKL094w	Predicted hydrolases and acyltransferases	Lipases (2LIP, 1TAH|B, 1CVL) •3 •8×10^{-5}	PldB is known to possess triglyceride lipase activity (41). All other proteins in the COG have not been characterized but now can be predicted to possess the α- or β-hydrolase fold.
e__cm_ COG0068C	HypF-sll0322-MJ0713	Hydrogenase maturation factor	Acylphosphatase (1APS) •2 •2×10^{-5}	HypF is required for hydrogenase biosynthesis (42), but no biochemical activity is known. The ~100 amino acid, NH_2-terminal domain aligns with acylphosphatase, with the catalytic residues conserved, suggesting that HypF orthologs indeed possess acylphosphatase activity. A PSI-BLAST search with this domain as the query detected five additional likely acylphosphatases, namely *E. coli* YccX and *M. jannaschii* MJ0809, MJ0553, MJ1331, and MJ1405 (43).
e__cm_ COG0663R	CalE, YrdA, YdbZ-sll1636, sll1031-MJ0304	Predicted carbonic anhydrases	Carbonic anhydrase from Methanosarcina thermophila (1THJ) •3 •10^{-29}	The biochemical activity of the proteins in this COG is not known. They show not only conservation of histidine residue comprising the active center of this unusual carbonic anhydrase (44) but also significant similarity to acetyltransferases of the isoleucine patch superfamily (45), suggesting an unexpected connection between the two types of enzymes.

*The designations are as in Table 1 and Fig. 3. †2g indicates two proteins from *M. genitalium*, 2p indicates two proteins from *M. pneumoniae*, and so forth. ‡The PDB accession is indicated in parentheses.

COGs as a crude first version of such a framework. Inclusion of additional, phylogenetically diverse genomes and further development of the procedures used to derive and analyze COGs will hopefully result in refinement of this system, making it a solid platform for genome annotation and evolutionary genomics.

REFERENCES AND NOTES

1. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
2. C. M. Fraser *et al.*, *ibid.* 270, 397 (1995); R. Himmelreich *et al.*, *Nucleic Acids Res.* 24, 4420 (1996); T. Kaneko *et al.*, *DNA Res.* 3, 109 (1996); F. R. Blattner *et al.*, *Science* 277, 1453 (1997).
3. C. J. Bult *et al.*, *Science* 273, 1058 (1996).
4. A. Goffeau *et al.*, *ibid.* 274, 546 (1996); H. W. Mewes *et al.*, *Nature* 387, 7 (1997).
5. C. R. Woese, *Curr. Biol.* 6, 1060 (1996); G. J. Olsen and C. R. Woese, *Cell* 89, 991 (1997); E. V. Koonin, *Genome Res.* 7, 418 (1997).
6. E. V. Koonin, A. R. Mushegian, K. E. Rudd, *Curr. Biol.* 6, 404 (1996); E. V. Koonin and A. R. Mushegian, *Curr. Opin. Genet. Dev.* 6, 757 (1996).
7. W. M. Fitch, *Syst. Zool.* 19, 99 (1970). This definition may not embrace all of the complexity of relationships between genes in different genomes. For example, if genes A and B are paralogs encoded in genome 1, and A′ and B′ are their respective orthologs in genome 2, what is the appropriate description of the relationship between A and B′? They formally are not paralogs, even though a generalized definition might include such cases. Furthermore, one-to-many and many-to-many orthologous relationships evidently exist.
8. W. M. Fitch, *Philos. Trans. R. Soc. London Ser. B* 349, 93 (1995).
9. R. L. Tatusov *et al.*, *Curr. Biol.* 6, 279 (1996).
10. E. V. Koonin, A. R. Mushegian, M. Y. Galperin, D. R. Walker, *Mol. Microbiol.* 25, 619 (1997).
11. The protein sequences were from the original references (*1*–*4*), with modifications (for example, tentative correction of frame-shift errors) and additions (previously unreported predicted genes) made for *E. coli* (E. V. Koonin and R. L. Tatusov, unpublished observations; K. E. Rudd, personal communication), *H. influenzae* (*9*), *M. genitalium* and *M. jannaschii* (*10*), and *S. cerevisiae* (T. J. Wolfsberg and D. Landsman, personal communication). The list of systematic names for all *E. coli* genes was provided by K. Rudd, and the names for all yeast genes were provided by T. Wolfsberg and D. Landsman; the *H. influenzae* genes were renamed as previously described (*9*); the gene names for the other species were from the original publications. The resulting protein database from complete genomes used in all comparisons contained 4283 sequences from *E. coli*, 1703 sequences from *H. influenzae*, 468 sequences from *M. genitalium*, 677 sequences from *M. pneumoniae*, 3168 sequences from *Synechocystis* sp., 1736 sequences from *M. jannaschii*, and 5932 sequences from *S. cerevisiae*, totaling 17,967 sequences. This sequence set is available on the World Wide Web at http://www.ncbi.nlm.nih.gov/COG. All pairwise comparisons between these sequences were performed using the BLASTPGP program, which is based on an enhanced version of the BLAST algorithm and includes analysis of local alignments with gaps (*26*). Predicted coiled coil regions in protein sequences were masked before the comparison using the batch version of the COILS2 program [A. Lupas, *Methods Enzymol.* 266, 513 (1996); D. R. Walker and E. V. Koonin, *ISMB* 5, 333 (1997)], and additionally, regions of low complexity were masked using the SEG program with default parameters [J. C. Wootton and S. Federhen, *Methods Enzymol.* 266, 554 (1996)]. Before the detection of triangles of BeTs, paralogs were identified as those proteins from the same lineage that showed greater similarity to each other than to any protein from another lineage. For the purpose of triangle formation, paralogs were treated as a group. The algorithm further included verification that the BeTs included in a triangle formed a consistent multiple alignment; triangles that did not contain a conserved motif were disregarded.
12. Although the exact solution depends on the amino acid composition and size of the particular proteins, under zero approximation, if B (from genome *b*) is the BeT for A (from genome *a*), and C (from genome *c*) is the BeT for B, the probability that C is the BeT for A by chance is close to $1/N$, where N is the number of genes in genome *c*, or ~0.001.
13. C. R. Woese, *Microbiol. Rev.* 51, 221 (1987); ———, R. Overbeek, G. J. Olsen, *J. Bacteriol.* 176, 1 (1994); N. R. Pace, *Science* 276, 734 (1997). A BeT to a given clade was registered if detected in any of the constituent species, for example, in *E. coli* or *H. influenzae* for the Gram-negative bacteria.
14. H. Watanabe and J. Otsuka, *Comput. Appl. Biosci.* 11, 159 (1995); E. V. Koonin, R. L. Tatusov, K. E. Rudd, *Methods Enzymol.* 266, 295 (1996).
15. A schematic visual representation of the search results was used for this analysis [T. L. Madden, R. L. Tatusov, J. Zhang, *Methods Enzymol.* 266, 131 (1996)].
16. A single-linkage clustering procedure was used with random match probability, $P < 0.001$, as the cutoff (*14*).
17. A searchable database of COGs is available at http://www.ncbi.nlm.nih.gov/COG. Each COG was assigned a unique identification number, which includes a letter for the functional category (*19*) and a number (see examples in Fig. 1 and Tables 1 and 2).
18. M. Lonetto, M. Gribskov, C. A. Gross, *J. Bacteriol.* 174, 3843 (1992).
19. The broad functional categories of proteins were as defined previously (*9*), except that transcription was separated from replication, recombination, and repair. This classification is a modification of the system originally developed for *E. coli* proteins [M. Riley, *Microbiol. Rev.* 57, 862 (1993)].
20. A partially similar representation of some of the protein families from complete genomes has been recently published [R. A. Clayton, O. White, K. A. Ketchum, J. C. Venter, *Nature* 387, 459 (1997)].
21. R. F. Doolittle, D.-F. Feng, S. Tsang, G. Chao, E. Little, *Science* 271, 470 (1996).
22. R. Himmelreich *et al.*, *Nucleic Acids Res.* 25, 701 (1997).
23. A. R. Mushegian and E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* 93, 10268 (1996).
24. E. V. Koonin, A. R. Mushegian, P. Bork, *Trends Genet.* 12, 334 (1996).
25. J. N. Weiser, M. Shchepetov, S. T. Chong, *Infect. Immun.* 65, 943 (1997).
26. J. P. Gogarten *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 86, 6661 (1989); N. Iwabe *et al.*, *ibid.*, p. 9355; J. P. Gogarten, E. Hilario, L. Olendzewski, in *Evolution of Microbial Life*, D. McL. Roberts, P. Sharp, G. Alderson, M. Collins, Eds. (Cambridge Univ. Press, Cambridge, 1996), pp. 257–292.
27. S. F. Altschul *et al.*, *Nucleic Acids Res.* 25, 3389 (1997). The probability of a random match, $P < 0.001$, was used in all PSI-BLAST searches.
28. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, *EMBO J.* 1, 945 (1982); A. E. Gorbalenya and E. V. Koonin, *Nucleic Acids Res.* 17, 8413 (1989); M. Saraste, P. R. Sibbald, A. Wittinghofer, *Trends Biochem. Sci.* 15, 430 (1990).
29. Protein sequences can be submitted for searching against COGs at http://www.ncbi.nlm.nih.gov/COG/cognitor.html
30. F. C. Bernstein *et al.*, *J. Mol. Biol.* 112, 535 (1977).
31. G. Chanfreau, S. M. Noble, C. Guthrie, *Science* 274, 1511 (1996); A. Jenny, L. Minvielle-Sebastia, P. J. Preker, W. Keller, *ibid.* 274, 1514 (1996); G. Stumpf and H. Domdey, *ibid.*, p. 1517.
32. J.-F. Tomb *et al.*, *Nature* 388, 539 (1997).
33. E. V. Koonin, R. L. Tatusov, M. Y. Galperin, M. N. Rozanov, unpublished observations.
34. P. Green *et al.*, *Science* 259, 1711 (1993).
35. J. Neuhard and R. A. Kelln, in Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, ed. 2, 1996), pp. 580–599.
36. E. C. C. Lin, *ibid.*, pp. 307–342.
37. T. W. Morris, K. E. Reed, J. E. Cronan Jr., *J. Bacteriol.* 177, 1 (1995).
38. P. Bork, N. P. Brown, H. Hegyi, J. Schultz, *Protein Sci.* 5, 1421 (1996).
39. D. Richter, E. Niegemann, M. Brendel, *Mol. Gen. Genet.* 231, 194 (1992); R. Wolter, W. Siede, M. Brendel, *ibid.* 250, 162 (1996).
40. H. Hama, T. Kayahara, W. Ogawa, M. Tsuda, T. Tsuchiya, *J. Biochem.* 115, 1135 (1994).
41. T. Kobayashi *et al.*, *ibid.* 98, 101 (1985).
42. A. Colbeau *et al.*, *Mol. Microbiol.* 8,15 (1993).
43. M. N. Rozanov and E. V. Koonin, unpublished observations.
44. B. E. Alber and J. G. Ferry, *Proc. Natl. Acad. Sci. U.S.A.* 91, 6909 (1994); C. Kisker *et al.*, *EMBO J.* 15, 2323 (1996).
45. E. V. Koonin, *Protein Sci.* 4, 1608 (1995); M. N. Rozanov and E. V. Koonin, unpublished observations.

46. We thank A. Schaffer for modifying the PSI-BLAST program; R. Walker, H. Watanabe, and M. Rozanov for valuable help with date analysis; K. Rudd, T. Wolfsberg, and D. Landsman for unpublished data; and P. Bork, M. Galperin, M. Gelfand, A. Mushegian, P. Pevzner, M. Roytberg, M. Rozanov, and R. Walker for helpful discussions.